DOCUMENT RESUME

ED 401 309 TM 025 870

AUTHOR Myford, Carol M.; And Others
TITLE Constructing Scoring Rubrics: Using "Facets" To Study Design Features of Descriptive Graphic Rating Scales.
PUB DATE Apr 96
NOTE 61p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Evaluators; *Rating Scales; *Scoring; *Student Evaluation; *Test Construction; Test Use; Visual Arts
IDENTIFIERS *FACETS Computer Program; FACETS Model; National Assessment of Educational Progress; *Scoring Rubrics

ABSTRACT

Developing scoring rubrics to evaluate student work was studied, concentrating on the use of intermediate points in rating scales. How scales that allow for intermediate points between defined categories should be constructed and used was explored. In the recent National Assessment of Educational Progress (NAEP) visual arts field test, researchers experimented with several formats for constructing scoring rubrics. Some descriptive graphic rating scales (continuous score scales) were pilot tested by 11 raters who scored the NAEP visual arts test for grades 4 and 8. Descriptive graphic ratings were designed to evaluate 4 test production blocks from the assessment, for a total of 50 pieces of student work. The "Facets" computer software was used to analyze the rating data. Raters were able to use the descriptive rating scales reliably. Some of the constructed scales were able to support 7 to 10 rating points rather than the traditional 3 or 4 points. However, there was little appreciable gain in reliability for scales having more than five points. The particular features of the scale (such as defined midpoint) were not as important as the knowledge, skills, and motivation of the rater. An appendix contains the graphic rating scales. (Contains 2 figures, 11 tables, and 32 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 401 309

Constructing Scoring Rubrics: Using *Facets* to Study Design Features of Descriptive Graphic Rating Scales

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

CAROL M. MYFORD

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Carol M. Myford
Eugene Johnson
Ray Wilkins
Hilary Persky
Mary Michaels

BEST COPY AVAILABLE

Educational Testing Service
Princeton, NJ

Paper presented at the annual meeting of the American Educational Research Association, April 1996, New York

TM025870

How should one go about the task of devising scoring rubrics to evaluate students' work? What format should those rubrics take? While there are resources available that provide practical advice for constructing rubrics (e.g., Airasian, 1991; Herman, Aschbacher & Winters, 1992; Linn & Gronlund, 1995; Stiggins, 1987), such resources do not address the issue of whether certain rubric formats are psychometrically superior to others. In several large-scale performance assessment programs (i.e., Vermont, Kentucky, and California statewide assessment programs; Pittsburgh Arts PROPEL), 3- or 4-point rubrics are commonly used. When constructing these rubrics, assessment developers typically identify specific observable aspects of the performance and/or product that are to be evaluated (i.e., the performance criteria). They then devise rating scales for the criteria, each scale containing three or four categories. The qualities or characteristics of a response associated with each category are defined in narrative form as precisely as possible. Raters use these narrative descriptions of qualities or characteristics to decide which rating to assign.

A thorny problem inevitably arises when raters use such rubrics: What rating should a rater give if a student's work falls "in the cracks" between the defined categories of the scale? For example, let's suppose the rater judges the student's work to be clearly better than the qualities or characteristics of performance described as a "level 1" response but not as good as the qualities or characteristics described as a "level 2" response. How should the rater handle this situation when it arises? In their recent review of the statistical procedures used in the California Learning Assessment System, Cronbach, Bradburn, and Horvitz (1994) suggest that raters be encouraged to use intermediate scale values for students' "borderline" responses. In their view, if raters were allowed to use intermediate values, the accuracy of the rating process could be improved, thus reducing a source of measurement error. They recommend that, at the very least, raters should be encouraged to use midpoints between the defined categories (i.e., for a scale with three defined categories labeled 1, 2, and 3, the raters should also be allowed to use the intermediate points 1.5 and 2.5 if they so desired). As a further refinement, they advocate allowing not only the use of midpoints but also points on either side of the midpoint (i.e., not only 2.5 but also 2.4 if the rater were leaning more toward a 2 than a 3, and 2.6 if the rater were leaning more toward a 3 than a 2).

If we decide to follow the advice of Cronbach et al. and allow for intermediate points in our rating scales, there are some important decisions we must make: What formats should we use when we construct these scales? Do some rating formats have better track records than others? How many intermediate points should our scales have? Is there an optimum number? How can we tell whether our scales have too few, too many, or the right number of points? We turn to a brief review of the literature on rating formats and the literature on the relationship between number of response categories and reliability to help us answer these questions.

Comparative Studies of Rating Formats. Surprisingly few studies comparing rating formats have been carried out in education-related settings. However, this has been a focus of much research and heated debate in personnel/organizational psychology since the 1950's. There is quite an extensive literature of comparative studies of various rating formats (e.g., behaviorally anchored rating scales (BARS), forced-choice formats, mixed standard scales, graphic rating scales). Reviewers of this literature have generally agreed that no one format has emerged as clearly superior to the others. In Landy and Farr's (1983) words,

> After more than 30 years of serious research, it seems that little progress has been made in developing an efficient and psychometrically sound alternative to the traditional graphic rating scale. ... It appears likely that greater progress in understanding performance judgments will come from research on the rating process than from a continued search for the "Holy Format." (p. 90)

As Guion (1986) suggests, of more importance than the particular rating format used is the competence of the rater. Cronbach (1990) echoes similar sentiments: "Many reporting formats and scoring systems for ratings have been tried. On the whole, it appears that the knowledge and motivation of the informant affect validity more than do features of the scale" (p. 587).

Number of Response Categories and Reliability. After conducting a series of studies examining the effect of number of categories on reliability of ratings, Bendig (1952a, 1952b, 1953, 1954a, 1954b) concluded that the reliability of the scales he employed did not increase as the number of scale categories increased from 5 to 9. He found that reliability decreased for scales having 3 or fewer categories and for scales having 11 or more categories. When Finn (1972) examined the reliability of

the scales he used, he concluded that reliability dropped with fewer than 3 or with more than 7 categories. In their Monte Carlo studies of factors affecting rating reliability, Lissitz and Green (1975) and Jenkins and Taber (1977) agreed that there was little appreciable gain in reliability when the number of scale categories exceeded 5.

Based on their review of the findings from these studies, Landy and Farr (1983) recommended that rating scales not include more than 9 categories since "the weight of evidence suggests that individuals have limited capacities for dealing with simultaneous categories of heterogeneous information. This was suggested long ago by Miller (1956) in his now-famous 'seven, plus or minus two' dictum, and appears to generalize to rating behavior" (pp. 83-84). Somewhat tongue-in-cheek, Guion (1986) quips, "it sometimes seems as if the five-point scale has been decreed from heaven, but there are other options" (p. 349). Indeed, when one compares the recommendations of educational measurement experts, one finds substantial differences of opinion regarding optimal number of scale points. Linn and Gronlund (1995) recommend that scales use between 3 and 7 scale points, while Cronbach (1990) recommends 4- to 7-point scales. Mehrens and Lehmann (1991) suggest that a maximum of 10 points be used, but they believe that 5- to 7-point scales are appropriate for most purposes. By contrast, Payne (1992) suggests that the optimal number of categories is probably 7 to 9, but, depending on the nature of the task and sophistication level of the raters, scales having as many as 7 to 20 categories could appropriately be used, Payne asserts.

Background of the Project

In the recent National Assessment of Educational Progress (NAEP) visual arts field test, we experimented with several different formats for constructing scoring rubrics. Past NAEP assessments in other content areas have typically made use of rubrics laid out as 3- or 4-point rating scales. In scoring these assessments, it has not been uncommon for raters to encounter examples of students' works that are difficult to rate. Certain works often do not appear to fit into any of the given categories of a scale but, rather, seem to lie "in the cracks" between the defined categories. We decided to pilot some descriptive graphic rating scales (i.e., continuous score scales) as part of the NAEP visual arts field test so that we could

learn about how raters would use intermediate score points when they have that option.

A descriptive graphic rating scale has two defined endpoints. These points are connected by a horizontal line. Descriptive phrases identify different points along the continuum. For some scales, the descriptive phrases might be quite brief, while for other scales the phrases might be more extensive. When a rater uses the scale to evaluate a product or a performance, the rater makes a vertical slash along the line to indicate where along that continuum the work lies. Descriptive graphic rating scales can incorporate different design features (i.e., presence or absence of a defined midpoint, presence or absence of hatchmarks along the line that connects the endpoints). We were interested in learning about how raters used these design features and which features, if any, affected interrater reliability.

The descriptive graphic rating format seems particularly attractive for assessment in the arts because it emphasizes the continuous nature of many of the performance criteria in these fields. While some arts-related performance criteria can appropriately be defined in terms of discrete categories, a number of criteria central to these domains cannot. Additionally, as Popham (1990) points out, this format takes advantage of the fact that "many people can use visual images to help them make qualitative gradations in their ratings" (p. 297). Because the raters in this study were visual artists accustomed to representing images visually, we felt it was appropriate to try out this rating format with them.

Our research set out to provide answers to a number of questions we posed about descriptive graphic rating scales. Through our experimentation we hoped to learn about how raters employ these scales to make judgments about student work. We planned to use what we learned to help NAEP program personnel decide whether scales in this format might be used to score some of the production tasks included in the 1997 NAEP visual arts operational assessment. The specific questions we posed are listed below:

- How many categories do the descriptive graphic rating scales we designed support? How should we think about them? as 3-point scales? as 4-point scales? as 5-point scales? etc.

- What is the effect of the number of categories on reliability of ratings for these scales? Does reliability cease to increase and instead begin to decrease at a certain point? Is there a point beyond which there is little utility to be gained in adding scale categories?

- How do various design features of these scales affect reliability? Do raters use descriptive graphic rating scales with defined midpoints any more reliably than scales without defined midpoints? Do raters use descriptive graphic ratings scales with hatchmarks any more reliably than scales without hatchmarks? Are 5 hatchmarks any better than 3?

- Will raters use descriptive graphic rating scales reliably if in their training they are only shown and discuss examples of students' work (i.e., anchors) for the endpoints of the scale but not for the midpoint? If they see and talk about anchors that fall at both ends and in the middle, will they produce more reliable ratings? If they are shown anchors that fall at various points along the full continuum, will they produce even more reliable ratings? What's the "bare minimum" raters need in the way of training anchors in order to use these descriptive graphic rating scales reliably?

Method

Participants

Raters. Eleven of the raters who scored the grades 4 and 8 NAEP visual arts field test participated in this study. Eight were white females, one was a black female, and two were white males. Five raters had master's degrees in art, one had a bachelor's degree in art, and five had degrees in fields other than visual arts. Two had experience teaching art in grades K-6, and three had college-level art teaching experience. All were practicing artists whose own work covered a variety of arts specialties (i.e., painting, drawing, sculpture, photography, computer art, video, filmmaking, design, printmaking, fiber art, mural design and execution, collage). None of the raters had any previous experience using descriptive graphic rating scales to evaluate students' works of art. While all the raters had taken part in the scoring of the field test, none of them had scored the blocks of art-making activities included in this study.

Trainers. Two persons participated in the study as trainers of the raters. One was a white female, and one was a white male. Both had previously served as trainers during the scoring of the NAEP visual arts field test for grades 4 and 8. Both

had advanced degrees in visual arts. One had experience teaching art in K-12 settings, and both had college-level art teaching experience. Neither of the trainers had any previous experience training raters to use descriptive graphic rating scales to evaluate students' works of art.

Procedure

Development of Descriptive Graphic Rating Scales. Part II of the NAEP visual arts field test included several production blocks. A *production block* contains multiple related exercises. The exercises frequently make use of the same stimulus material, and a combination of exercise formats are employed. When devising production blocks, efforts were made to integrate three artistic processes (i.e., creating, performing and responding) within a block. The exercises contained in a block are designed to engage students in activities typical of these three artistic processes. We selected four of these blocks to focus on in our study (Blocks R1VAX1, R123VAX6, R23VAX5, and R23VAX7). The specific art-making activities upon which we concentrated our scale development efforts are described below:

- Block R1VAX1: Students selected a type of animal and then drew a comfortable place (an environment) for the animal. They were directed to make use of near and far shapes (i.e., perspective) and shapes that overlap when drawing their animal's place. They were also instructed to use the space and shape of their drawing paper in ways that were best for depicting the animal's place.

- Block R123VAX6: Students drew an idea for a mural to show something that was important to the people in their community. They were instructed to use shapes, lines, colors, and forms that would capture the attention of people from far away. As they worked on their design, they were asked to think about how they were using the drawing space.

- Block R23VAX5: Students created a self-portrait. They were instructed to use materials in a way that would communicate to a viewer something that they thought was important about their personality.

- Block R23VAX7: Students read an ancient legend and then experimented with ways to visually express figures in the legend. They were asked to creatively combine the figures into a complete drawing, showing how they might interact. Students were reminded to choose media (drawing tools) from their materials packet that would best help them express their ideas most effectively.

We designed descriptive graphic rating scales to evaluate works of art students created in these blocks. (See Appendix A for copies of the scales we constructed.) The individual scales exhibited different combinations of design features. For some of the scales, we defined two endpoints of the scale and a midpoint; for other scales, we defined only the two endpoints. For some of the scales we constructed a horizontal line to connect the endpoints and then placed either three hatchmarks or five hatchmarks at specific points along the line to show key transition points along the continuum. Other scales had no hatchmarks along the horizontal line.

Selection of Student Work. We selected samples of works of art that students created for these four production blocks during the field test. For each of the blocks, we used the ratings given during the scoring of the field test to assist us in selecting 50 pieces of student work that would represent the full range of student ability exhibited. (In the scoring of the field test, raters used 3-point scales to evaluate these works. When pulling the 50 samples of student work for the study, we included some samples that received all 3's, some that received all 2's, some that received all 1's, and some that received mixtures of 3's, 2's, and 1's.) The trainers selected additional samples of student work to serve as anchors for rater training purposes and to include in sets for raters to use as practice.

Rater Training. During each training session, the trainer introduced the raters to the two scales the raters would be using during that session. The trainer defined each of the performance criteria, and then raters examined and discussed samples of student work in order to clarify the meaning of each criterion and the distinctions between the various points on the scale. The trainer presented examples of students works (i.e., anchors) and talked about the characteristics of each that should be considered when assigning a rating. After introducing each work and talking about its characteristics, the trainer would fasten the work to the wall showing its position along the linear continuum.[1]

---

[1] In some ways, the scales we devised might be thought of as product scales (Linn & Gronlund, 1995) since we purposely chose examples of students' work that represented various levels of quality and then visually displayed them so that raters could develop a sense of the continuum of quality they were likely to see when they carried out the actual scoring. Indeed, during the scoring sessions, if a rater had difficulty deciding where to place his or her slash along the horizontal line, the rater would frequently bring the work to the wall and compare it to the anchors that were displayed to determine where the student's work seemed to best "fit" along the continuum.

In this study, we experimented with several approaches to using anchors during training. In one training session, the trainer showed and discussed anchors for five points along the continuum (i.e., the two endpoints and three points in between). For other training sessions, the trainer showed and discussed anchors for both endpoints of each scale and for the midpoint. In still other training sessions, the trainer showed and discussed anchors for only the endpoints. We wanted to know whether raters needed to see and talk about anchors along the full continuum in order to use the scales reliably, or whether they could score reliably after having seen anchors for two endpoints and a midpoint, or for endpoints only.

After the trainers introduced and discussed the anchors, time was set aside for raters to practice scoring samples of student work. The raters would independently score small sets of five student works and then, as a group, discuss the ratings they gave. Using a flip chart, the trainer would draw five horizontal lines and ask each rater to come forward and indicate where along each line he or she had placed each student's work. After all the raters had taken turns making their slash marks, the group discussed their ratings in order to clarify meanings of each of the performance criteria and to attempt to reach consensus in their usage of the rating scales.

Scoring the Student Work. The scoring took place over a two-day period. The experimental design we employed is shown in Figures 1 and 2. We randomly

---

Insert Figures 1 and 2 about here

---

assigned the eleven raters to two groups. Group 1 met with Trainer 1 and completed training to score Block R123VAX6/AM. The raters then scored the 50 students' works selected for that block. Concurrently, Group 2 met with Trainer 2, completed training to score Block R1VAX1/AM, and then scored the 50 students' works selected for that block. Following a lunch break, Group 1 met with Trainer 2 and were trained to score Block R1VAX1/PM. The raters scored the same set of 50 students' works that Group 2 had scored in the morning, but the scales they used had different design features than the scales Group 2 had used (see Figure 1 for a description of those design features). Group 2 met with Trainer 1 to learn to score

Block R123VAX6/PM.[2] Group 2 scored the same set of 50 students' works that Group 1 had scored in the morning, but the scales they used had different design features than the scales Group 1 had used.

On the second day, we again randomly assigned the eleven raters to two new groups. Group 1 met with Trainer 1 and completed training to score Block R23VAX5/AM. The raters then scored the 50 students' works selected for that block. Concurrently, Group 2 met with Trainer 2, completed training to score Block R23VAX7/AM, and then scored the 50 students' works selected for that block. Following a lunch break, Group 1 met with Trainer 2 and were trained to score Block R23VAX7/PM. The raters scored the same set of 50 students' works that Group 2 had scored in the morning, but the scales they used had different design features than the scales Group 2 had used (see Figure 2 for a description of those design features). Group 2 met with Trainer 1 to learn to score Block R23VAX5/PM. Group 2 scored the same set of 50 students' works that Group 1 had scored in the morning, but the scales they used had different design features than the scales Group 1 had used.

At the end of the second day, we gave each rater a questionnaire to fill out to gather their reactions to using the experimental rating scales. We provided them with postage-paid envelopes and asked them to return the completed questionnaires within a week.

Data Analysis

To analyze the rating data from this study, we employed *Facets* (Linacre, 1994a), a Rasch-based rating scale analysis computer software program. *Facets* is a generalization of Wright and Masters' (1982) Partial Credit model which makes possible the analysis of data from assessments that have more than the traditional two "facets" associated with multiple-choice tests (i.e., "items" and "examinees").

In the many-facet Rasch model (Linacre, 1994b), each "element" of each facet of the assessment situation (e.g., each student, rater, rating scale category, etc.) is represented by one parameter. In this study, the model contains a parameter

[2]Note that each of the four blocks was scored twice--once in the morning, and once in the afternoon (by a different set of raters). Hence, the designation "AM" or "PM" appearing after each block.

representing student "ability," a second parameter representing rater "severity," and a third parameter representing rating scale category "challenge." *Facets* has the form of a log-linear model for main effects, and estimates those effects in "logits," or the logarithms of odds of a given rating compared to the next lower one. For our study, the model takes the following particular form: the log-odds of the probability that a student with a "true" ability of θ will receive from Rater j a rating in Category k [denoted $P_{j,k}(\theta)$] as opposed to receiving a rating in Category k-1 [denoted $P_{j,k-1}(\theta)$] on a rating scale with k categories is modeled as

$$\ell n\left[P_{j,k}(\theta)/P_{j,k-1}(\theta)\right] = \theta - \xi_j - \tau_k, \qquad (1)$$

where ξ_j is the "severity" parameter associated with Rater j, and τ_k for k=2,...,K is a parameter indicating the relative probability of a rating in Category k as opposed to Category k-1 for the scale when $\tau_1 \equiv 0$. It follows that the probability of a rating in category k for a student with parameter θ from Rater j is

$$P_{j,k}(\theta) = \frac{\exp\left[k(\theta - \xi_j) - \sum_{s=1}^{k}\tau_s\right]}{\sum_{t=1}^{K}\exp\left[t(\theta - \xi_j) - \sum_{s=1}^{t}\tau_s\right]} \qquad \text{for k = 1, K.} \qquad (2)$$

When raters evaluated students' work, they were instructed to place a vertical slash along an 8-1/2" horizontal line to indicate where along that continuum they felt the students' work fell. To prepare these data for analysis, we measured from the left end of each line to the point where the rater's slash crossed that line, rounding to the nearest 1/16". We converted that number to its decimal equivalent and then transformed this 0 to 8.5 scale to a scale that ran from 0 to 255 (i.e., the maximum number of rating scale categories *Facets* can accommodate is 255).

For each block, we ran a series of eight *Facets* analyses. For example, for Block R1VAX1/AM we first analyzed the data as if the two scales were 3-point scales and then examined how the rating scales functioned. (For this analysis, we divided the horizontal line (i.e., the 0 to 255 scale) into three equal segments: ratings from 1-85 were recoded as "1," ratings from 86-170 were recoded as "2," and ratings from 171-255 were recoded as "3.") Using *Facets* recoding capabilities, we then ran additional

analyses to see how the scales would function if we were to consider the scales as 4-point scales (i.e., dividing the 0 to 255 horizontal scale into four equal segments), as 5-point scales, as 6-point scales, as 7-point scales, as 8-point scales, as 9-point scales, and, finally, as 10-point scales.[3] By comparing the output from the various analyses, we sought to determine what the optimum rating scale structure was for each scale from the standpoint of measurement precision and scale discriminability.

In addition to the *Facets* analyses, we also ran intraclass correlational analyses (Berk, 1979; Cherry & Meyer, 1993; Cronbach, Ikeda, & Avener, 1964; Ebel, 1951; Shrout & Fleiss, 1979) so that we could compare the Rasch student separation reliabilities reported as part of the *Facets* output to conventional intraclass correlations. Based on analysis of variance procedures, intraclass correlation expresses the "classical theory of measurement error relationship between true and observed variance" (Berk, p. 463). In our study, all raters rated all students included in each block. Therefore, we used the following formula for fully crossed designs as recommended by Cherry and Meyer (1993) to calculate intraclass correlation:[4]

$$r = \frac{MS_p - MS_e}{MS_p + (k-1)MS_e} \tag{3}$$

where MS_p is the between-persons mean square, MS_e is the error mean square, and k is the number of raters.

---

[3] An alternative strategy for recoding the ratings involves defining the rating scale categories such that the categories have nearly equal numbers of ratings (i.e., counts) in each (J. M. Linacre, personal communication, Nov. 5, 1995). For example, if 6 raters each rated 50 pieces of student work on a single descriptive graphic rating scale, 300 ratings would be generated. Suppose we wanted to analyze these ratings to see how the scale would function as a 3-point scale using the *equal counts* recoding strategy. If 100 of those ratings fell between 1 and 115, we would recode each of these ratings to "1." If the next 100 ratings fell between 116 and 145, we would recode these ratings as "2." If the last 100 ratings fell between 146 and 255, we would recode these ratings as "3." Using this recoding strategy, we would now have a 3-point scale with the three categories each containing an equal numbers of ratings. Note how the *equal counts* recoding strategy differs from the recoding strategy we used in this study (i.e., defining the rating scale categories by dividing the horizontal line--the 0 to 255 scale--into *equal segments* rather than dividing the total number of ratings given into *equal counts*. Initially, we ran a series of *Facets* analyses using the *equal counts* recoding strategy and another set of *Facets* analyses on the same data (i.e., 4 of the 8 blocks) using the *equal segments* recoding strategy so that we could compare output from both sets of analyses. In each case, there was very little difference between the two sets of output when we examined key indicators (i.e., student separation, rater separation, interrater reliability attenuated by rater variance). Therefore, we decided to include in this report only findings from the analyses in which we used the *equal segments* recoding strategy.

[4] This formula is based on a two-way mixed effects ANOVA having two independent variables--"raters" (which is treated as a fixed effect) and "students" (which is treated as a random effect). Cherry and Meyer (1993) describe how between-rater variance is treated when computing intraclass correlation using this formula: "The difference in average scores between two raters (or among three or more raters) is attributed to raters consistently applying slightly different standards of judgment rather than to error, and the difference in average scores is therefore considered true variance" (p. 133).

Results

Our study sought to answer four related sets of questions. We structured our discussion of research findings around the specific questions we explored with the *Facets* output.

- *How many categories do the descriptive graphic rating scales we designed support? How should we think about them? as 3-point scales? as 4-point scales? as 5-point scales? etc.*

Facets provides several pieces of output that can help us answer these questions. For each scale, *Facets* reports the percentage of ratings that fall into each category which facilitates examination of category usage by raters. Beyond this, *Facets* reports the "Average Measure Difference" (AMD) for each category on a scale. Linacre (1994b) defines *average measure difference* as "the average of the [student] measures [of ability] that are modeled to generate the observations in this category" (p. 69). As we move from categories at the lower end of a scale to categories at the higher end of a scale, we would hope to see a pattern of ascending AMD's (Linacre, 1995). When we see evidence of this kind of pattern occurring, it suggests that the rating scale categories are appropriately ordered and are functioning properly. Higher ratings do correspond to "more" of the variable being rated. If AMD's do not increase (i.e., if we see identical values for adjacent categories or one or more descending values), then that suggests that some of the categories are not functioning as intended. (For example, while we may have intended for a scale we designed to function as a 5-point scale, the raters may instead be using it as a 3- or 4-point scale. Raters may find that some of the categories are not clearly differentiated from one another.)

Facets provides an additional check on category ordering. For each category, *Facets* reports "the lowest [student ability] measure at which this category is the one most probable to be observed" (Linacre, 1994b). *Facets* identifies those categories that are never most probable to be observed for *any* student ability measure. Like the AMD's, these "Most Probable" Thresholds (MPT's) should also be ordered, increasing as we move from categories at the lower end of a scale to categories at the higher end of a scale. If a category is never most probable to be observed, then that suggests that there are problems with the rating scale categories (i.e., some categories

are not distinguishable and are underutilized) and may signal a need to reduce the number of categories by combining some of them. As Andrich (1996) notes, a scale may show ascending AMD's but not ascending MPT's.

In Tables 1 through 8 we report both AMD's and MPT's for the descriptive graphic scales used in this study. For each scale, we show how many ascending values appeared in the output for the AMD's and for the MPT's. When we look across values reported for Scales 1 and 2 in the rows labeled "Most Probable From" and "Average Measure Difference" within a table, we can get a sense of how many categories each of the two descriptive graphic rating scales in that block could support. For example, in Table 1 we note that when we analyzed Scale 1 as a 3-point scale, the three categories on that scale had ascending AMD's and ascending MPT's. It would be appropriate, then, to think of Scale 1 as a 3-point scale. As we look across the Scale 1 "Most Probable from" row and the Scale 1 "Average Measure Difference" row, we see that Scale 1 would also support an interpretation of it as a 4-point, 5-point, 6-point, or 7-point scale. In each case, all the AMD's and MPT's are ascending for Scale 1. We see, though, that when we analyzed Scale 1 as a 8-point scale, the 8 categories on that scale had ascending MPT's, but only 7 categories had ascending AMD's. These findings would cast some doubt on whether Scale 1 would support an 8-point interpretation. When we analyzed Scale 1 as a 9-point scale, only 8 categories had ascending MPT's. It appears, then, that if we use "Average Measure Differences" as the decision-making criterion, the number of scale points that Scale 1 would support is 7. By contrast, if we were to use the "Most Probable" Thresholds as the decision-making criterion, the number of scale points that Scale 1 would support is 8.

Insert Tables 1 to 8 about here

What is the number of scale points that our scales could support? When we examine summary Table 9, we see that if we were to use "Most Probable" Thresholds as our decision-making criterion, then we would conclude that all the scales could support at least a 5-point interpretation, while some of the scales could be thought of as supporting as many as 7 or 8 points. However, if we were to use "Average Measure Differences" as our decision-making criterion, then we would

conclude that all the scales could support at least a 7-point interpretation, while some of the scales could be thought of as supporting as many as 9 or 10 points.

Insert Table 9 about here

- *What is the effect of the number of categories on reliability of ratings for these scales? Does reliability cease to increase and instead begin to decrease at a certain point? Is there a point beyond which there is little utility to be gained in adding scale categories?*

To answer these questions, we focus on student separation reliabilities and intraclass correlation coefficients contained in Tables 1 through 8. In Rasch terms, *student separation* is a measure of the spread of the estimates of student ability relative to their precision (Linacre, 1994b). The student separation index indicates the number of statistically different strata of student ability in the sample of students evaluated by the rating scales (Wright, 1996). Student separation has a range of 0 to ∞. When we look across Tables 1 to 8, it appears that we could consistently identify between 2 and 4 student strata, depending upon the number of points on the scales used. Generally, as the number of rating scale points increases, student separation increases. However, for each block, there is a certain point at which the amount of increase levels off, or, in some cases, actually begins to decrease.

While *Facets* does not provide a measure of interrater reliability *per se*, it does include a separation reliability which, like interrater reliability, has a range of 0 to 1. The Rasch student separation reliability is the ratio of "true" variance in student scores to the "observed" variance in student scores. As Wright explains (1996), "In Rasch terms, 'true' variance is the 'adjusted' variance (observed variance adjusted for measurement error). Error variance is a mean-square error (derived from the model) inflated by misfit to the model encountered in the data" (p. 472). The student separation reliabilities we report in Tables 1 through 8 have been attenuated for rater variance (Linacre, 1991). As Linacre (1991) notes, there is usually little difference between Rasch student separation reliabilities and interrater reliabilities when equivalent variance terms are used to compute them. Student separation reliabilities for the eight blocks are generally in the range of .70 to .90. As the number of rating scale points increases, separation reliability increases; but for each

block, we reach a point of diminishing returns (i.e., the reliability ceases to increase, or, in some cases, actually decreases). In some blocks, that leveling off or decrease tends to occur as we move from 5-point scales to 6-point scales (for blocks R1VAX1/AM, R1VAX1/PM, R123VAX6/AM, R23VAX7/AM, R23VAX7/PM), while for other blocks this occurs as we move from 7-point scales to 8-point scales (for blocks R123VAX6/PM, R23VAX5/AM, and R23VAX5/PM).

To facilitate comparison of Rasch student separation reliability with a more traditional measure of interrater reliability, we calculated intraclass correlation coefficients. For most of the blocks, the intraclass correlations tend to be somewhat lower than the comparable student separation reliabilities (although for three of the blocks--R23VAX7/PM, R23VAX5/AM, and R23VAX5/PM--the intraclass correlations are actually somewhat higher than the separation reliabilities). For all the blocks, the intraclass correlations are generally in the range of .65 to .90. As the number of rating scale points increases, intraclass correlation increases. However, we reach a point of diminishing returns for each block (i.e., the correlations cease to increase, or, in some cases, begin to decrease), just as occurred with the student separation reliabilities. That leveling off (or decrease) tends to occur as we move from 4-point scales to 5-point scales for block R23VAX7/PM; from 5-point scales to 6-point scales for blocks R123VAX6/PM, R23VAX7/AM, and R23VAX5/AM; from 6-point scales to 7-point scales for blocks R123VAX6/AM and R23VAX5/PM; and from 7-point scales to 8-point scales for blocks R1VAX1/AM and R1VAX1/PM.

To summarize, it appears that for both student separation and intraclass correlation little appreciable gain in reliability occurs if we think about these scales as having more than 5 points. In general, moving from 3-point scales to 5-point scales results in a useful gain in reliability (i.e., ranging from a gain of .03 to .10 for student separation, and a gain of .03 to .14 for intraclass correlation). However, moving from 5-point scales to 10-point scales nets, at best, a .03 gain for student separation (for blocks R123VAX6/PM and R23VAX7/PM) and a .04 gain for intraclass correlation (for block R23VAX7/PM). (More often, as we move from thinking about these as 5-point scales to thinking about them as 10-point scales, the gain in reliability is on the order of .01 to .02 for both these indices.)

- *How do various design features of these scales affect reliability? Do raters use descriptive graphic rating scales with defined midpoints any more reliably*

than scales without defined midpoints? Do raters use descriptive graphic ratings scales with hatchmarks any more reliably than scales without hatchmarks? Are 5 hatchmarks any better than 3?

To answer these questions, we compared the student separation reliabilities (Table 10) and the intraclass correlation coefficients (Table 11) for the eight blocks for 5-point scales. (We used the 5-point scale data since all of the scales included in this study could support a 5-point interpretation and, for the most part, there seemed to be little appreciable gain in reliability beyond 5 points.) In each table, we ordered the indices (i.e., the separation reliability coefficients and the intraclass coefficients) from high to low. For each block we included information about the design features of the scales in that block (i.e., presence or absence of a defined midpoint and number of hatchmarks).

Insert Tables 10 and 11 about here

The student separation reliabilities reported in Table 11 range from .79 to .91. Whether or not a scale had a defined midpoint did not seem to affect separation reliability. When we examine the blocks with the *highest* separation reliabilities, we see that some of the blocks had a defined midpoint (i.e., R1VAX1/PM), while others did not (i.e., R123VAX6/PM). Similarly, when we examine the blocks with the *lowest* separation reliabilities, we see that some of the blocks had a defined midpoint (i.e., R23VAX7/PM), while others did not (i.e., R23VAX5/AM). Also, there does not appear to be a consistent relationship between number of hatchmarks and separation reliability. The blocks with the *highest* separation reliabilities contain no hatchmarks (i.e., R1VAX1/PM and R123VAX6/PM), but the block with the *lowest* separation reliability also contained no hatchmarks (i.e., R23VAX7/PM).

The intraclass correlation coefficients reported in Table 11 range from .75 to .91. Whether or not a scale had a defined midpoint did not seem to affect intraclass correlation. When we examine the blocks with the *highest* intraclass correlations, we note that some of the blocks had a defined midpoint (i.e., R23VAX5/PM), while others did not (i.e., R23VAX5/AM). Similarly, when we examine the blocks with the *lowest* intraclass correlations, we see that some of the blocks had a defined midpoint (R123VAX6/AM), while others did not (i.e., R1VAX1/AM and

R123VAX6/PM). Additionally, it does not appear that number of hatchmarks affects intraclass correlation. When we examine the two blocks containing scales with 5 hatchmarks, we see that one of the blocks had the *highest* intraclass correlation (i.e., R23VAX5/PM), while the other block had one of the *lowest* intraclass correlations (i.e., R1VAX1/AM).

- *Will raters use descriptive graphic rating scales reliably if in their training they are only shown and discuss examples of students' work (i.e., anchors) for the endpoints of the scale but not for the midpoint? If they see and talk about anchors that fall at both ends and in the middle, will they produce more reliable ratings? If they are shown anchors that fall at various points along the full continuum, will they produce even more reliable ratings? What's the "bare minimum" raters need in the way of training anchors in order to use these descriptive graphic rating scales reliably?*

To answer these questions we refer to Tables 10 and 11. Each table includes information about the anchors and practice sets used in training for each block. When we review these tables, we see little evidence of a consistent relationship between the nature of the anchors used in training and reliability.

Table 10 reveals that some of the blocks with the *highest* separation reliabilities had anchors showing endpoints and a midpoint (i.e., R1VAX1/PM), while other blocks had anchors showing endpoints only (i.e., R123VAX6/PM and R1VAX1/AM). Similarly, when we examine the blocks with the *lowest* separation reliabilities, we see that some of the blocks had anchors showing endpoints and a midpoint (i.e., R23VAX7/PM), while other blocks had anchors showing endpoints only (i.e., R23VAX5/AM). The block that had anchors showing all five scale points (i.e., R23VAX5/PM) had neither the highest nor the lowest separation reliability.

We see much the same story when we review the information in Table 11, with one difference: the block that had anchors showing all five scale points (i.e., R23VAX5/PM) had the *highest* intraclass correlation (.91). The three blocks having anchors that showed endpoints and a midpoint (i.e., R1VAX1/PM, R23VAX7/PM and R123VAX6/AM) had intraclass correlations in the range of .80 to .85. We see somewhat greater variation across the blocks having anchors that showed only endpoints. Some of these blocks (i.e., R23VAX5/AM and R23VAX7/AM) had intraclass correlations in the range of .87 to .91, while other blocks (i.e.,

R1VAX1/AM and R123VAX6/PM) had lower intraclass correlations in the range of .75 to .79.

Finally, it is interesting to note that the one block that had practice sets showing examples of student work for endpoints only (i.e., R123VAX6/PM) had the *lowest* intraclass correlation (.75) but one of the *highest* separation reliabilities (.90).

Discussion

Can trained raters use descriptive graphic rating scales to evaluate students' works of art? Based on findings from our study, we conclude that the raters were able to reliably use the scales we constructed. We found that all the scales would support at least a 5-point interpretation, and that individually some of the scales could be thought of as supporting as many as 7 to 10 points. These findings lend support to the suggestion made by Cronbach et al. (1994) that raters be encouraged to use midpoints between defined categories to improve the accuracy of the rating process, thus reducing a source of measurement error. It appears that the raters in this study were able to make finer distinctions than a traditional 3- or 4-point scoring rubric allows. However, it's important to note that we found little appreciable gain in reliability for scales having more than 5 points, confirming the findings of Bendig (1952a, 1952b, 1953, 1954a, 1954b), Lissitz and Green (1975), and Jenkins and Taber (1977). In general, moving from 3-point to 5-point scales resulted in a useful gain in reliability, but the net gain in reliability associated with moving from 5-point to 10-point scales was minimal.

When we designed the descriptive graphic rating scales for this study, we varied certain design features of the scales. The individual scales exhibited different combinations of the design features. For some of the scales, we defined two endpoints of the scale and a midpoint; for other scales, we defined only the two endpoints. For some of the scales we constructed a horizontal line to connect the endpoints and then placed either three or five hatchmarks at specific points along the line to show key transition points along the continuum. Other scales had no hatchmarks along the horizontal line. We wanted to determine whether these two design features (i.e., presence or absence of a defined midpoint, number of hatchmarks) affected interrater reliability. We looked at two measures of rater reliability: Rasch student separation reliabilities and intraclass correlations. We

computed these measures for our scales, considering the scales as 5-point scales. The student separation reliability coefficients for the 5-point scales ranged from .79 to .91. The intraclass correlation coefficients for these same scales ranged from .75 to .91. Whether or not a scale had a defined midpoint did not affect student separation or intraclass correlation. Similarly, whether the scale had 0, 3, or 5 hatchmarks did not affect these measures. These findings lend credence to the views of Guion (1986) and Cronbach (1990) who contend that the particular features of a scale are not as important as the knowledge, skills, and motivation of the rater.

We experimented with several approaches to using anchors during training. In one training session, the trainer showed and discussed anchors for five points along the continuum (i.e., the two endpoints and three points in between) and then provided practice sets for the rater to use that contained students' works covering the full continuum. For other training sessions, the trainer showed and discussed anchors for both endpoints of each scale and for the midpoint, and then raters practiced scoring works that covered the full continuum. In still other training sessions, the trainer showed and discussed anchors for only the endpoints, and the raters then practiced scoring works covering the full continuum. We wanted to know whether raters needed to see and talk about anchors along the full continuum in order to use the scales reliably, or whether they could score reliably after having seen anchors for two endpoints and a midpoint, or for endpoints only. Our findings would suggest that raters can use descriptive graphic rating scales reliably if they see and talk about anchors for both endpoints of the scale but then have some practice rating examples of students' works that cover the full continuum of student ability.

If NAEP program personnel were to decide to include some scales in the descriptive graphic format to score some of the production tasks included in the 1997 NAEP visual arts assessment, then there are some operational concerns that will need to be addressed. As a first step, we would need to consider ways to streamline the process of translating a rater's slash on a line into a score. For this study, we manually carried out the various steps in this process, measuring from the left end of each line to the point whether the rater's slash crossed that line, and then rounding to the nearest 1/16". After we converted that number to its decimal equivalent, we transformed the 0 to 8.5 scale to a scale that ran from 0 to 255. This was a laborious and time-consuming process that, perhaps, could be automated. We could investigate the feasibility of transferring the rating scales to a computer,

having the raters use a mouse to click at the point on the horizontal line where they judge a student's work to lie, and then having the computer convert that mark into a score. If we could use technology to automate the steps in this data preparation process, then that could result in considerable time and cost savings.

Traditionally, when raters use scoring rubrics in NAEP, program personnel overseeing the scoring process consider ratings that are more than 2 points apart (or, in the case of 3-point rubrics, more than 1 point apart) discrepant, and a third rater is brought in to adjudicate the discrepancy. But what does "discrepancy" mean for raters using descriptive graphic rating scales? How far apart do two raters' slash marks on a horizontal line need to be in order to be considered discrepant? Suppose we were to establish a guideline for defining what we mean by discrepancy. Could we then program a computer to identify students' works that received discrepant ratings so that they could be set aside for third-rater adjudication?

If we were to incorporate descriptive graphic rating scales into NAEP assessments, we would also need to work through a number of issues related to combining and reporting assessment results:

- Can we combine results from students' performance on production tasks scored using descriptive graphic rating scales alongside results from students' performance on other types of exercises included in the NAEP visual arts assessment (i.e., short constructed response exercises, multiple-choice items, extended constructed response exercises, production tasks scored using traditional 3- and 4-point rubrics)? Is it psychometrically feasible to produce a single unidimensional scale that will encompass these diverse sources of information about students' performance? If we cannot produce a single unidimensional scale, is it feasible to produce several scales?

- If we were to report narratively on student performance for production tasks scored using descriptive graphic rating scales, what type of reporting format should we use? (For NAEP assessments in other content areas, illustrative exercises are often presented as part of the final report, and the percentage of student responses falling into each category are reported. However, these illustrative exercises have typically been scored using rubrics containing 3 or 4

discrete categories. How would we report on student performance if our scales do not contain discrete categories?)

- How will the information we derive from scoring these production exercises using the descriptive graphic rating scales feed into achievement level reporting? Can we use this scoring information to help us define basic, proficient, and advanced achievement levels in the visual arts?

Conclusions

The descriptive graphic rating scale format seems to hold promise as a suitable format for scoring production tasks to be included in the 1997 NAEP visual arts assessment. The scales we piloted in this study seemed to work quite well and, according to the indicators we used, appear psychometrically sound. As next steps, there are issues related to combining and reporting assessment results that will need to be addressed before such scales could become part of a NAEP operational assessment.

References

Andrich, D. (1996). Category ordering and their utility. *Rasch Measurement: Transactions of the Rasch Measurement SIG, 9*(4), 464-465.

Bendig, A. W. (1952a). A statistical report on a revision of the Miami instructor rating sheet. *Journal of Educational Psychology, 43*, 423-429.

Bendig, A. W. (1952b). The use of student rating scales in the evaluation of instructors in introductory psychology. *Journal of Educational Psychology, 43*, 167-175.

Bendig, A. W. (1953). The reliability of self-ratings as a function of the amount of verbal anchoring and of the number of categories on the scale. *Journal of Applied Psychology, 37*, 38-41.

Bendig, A. W. (1954a). Reliability and number of rating scale categories. *Journal of Applied Psychology, 38*, 38-40.

Bendig, A. W. (1954b). Reliability of short rating scales and the heterogeneity of the rated stimuli. *Journal of Applied Psychology, 38*, 167-170.

Berk, R. A. (1979). Generalizability of behavioral observations: A clarification of interobserver agreement and interobserver reliability. *American Journal of Mental Deficiency, 83*, 460-472.

Cherry, R. D., & Meyer, P. R. (1993). Reliability issues in holistic assessment. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment* (pp. 109-141). Cresskill, NJ: Hampton Press.

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper & Row.

Cronbach, L. J., Bradburn, N. M., & Horvitz, D. G. (1994, July). *Sampling and statistical procedures used in the California Learning Assessment System. Report of the Select Committee.* Palo Alto, CA: Author.

Cronbach, L. J., Ikeda, M., & Avner, R. A. (1964). Intraclass correlation as an approximation to the coefficient of generalizability. *Psychological Reports, 15*, 727-736.

Ebel, R. L. (1951). Estimation of the reliability of ratings. *Psychometrika, 16*, 407-424.

Finn, R. H. (1972). Effects of some variations in rating scale characteristics on the means and reliabilities of ratings. *Educational and Psychological Measurement, 32*, 255-265.

Guion, R. M. (1986). Personnel evaluation. In R. A. Berk (Ed.), *Performance assessment: Methods and applications* (pp. 345-360). Baltimore: Johns Hopkins University Press.

Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). *A practical guide to alternative assessment.* Alexandria, VA: Association for Supervision and Curriculum Development.

Jenkins, G. D., & Taber, T. A. (1977). A Monte Carlo study of factors affecting three indices of composite scale reliability. *Journal of Applied Psychology, 62,* 392-398.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87,* 72-107.

Landy, F. J., & Farr, J. L. (1983). *The measurement of work performance: Methods, theory, and applications.* San Diego: Academic Press, Inc.

Linacre, J. M. (1991). Inter-rater reliability. *Rasch Measurement: Transactions of the Rasch Measurement SIG, 5*(3), 166.

Linacre, J. M. (1994a). *Facets* [Computer program]. Chicago, IL: MESA Press.

Linacre, J. M. (1994b). *A user's guide to Facets Rasch measurement computer program.* Chicago, IL: MESA Press.

Linacre, J. M. (1995). Categorical misfit statistics. *Rasch Measurement: Transactions of the Rasch Measurement SIG, 9*(3), 450-451.

Linn, R. L., & Gronlund, N. E. (1995). *Measurement and assessment in teaching* (7th ed.). Englewood Cliffs, NJ: Merrill.

Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: A Monte Carlo approach. *Journal of Applied Psychology, 60,* 10-13.

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology.* Chicago: Holt, Rinehart and Winston, Inc.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81-97.

Payne, D. A. (1992). *Measuring and evaluating educational outcomes.* New York: Macmillan.

Popham, W. J. (1990). *Modern educational measurement: A practitioner's perspective.* Englewood Cliffs, NJ: Prentice Hall.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420-428.

Stiggins, R. J. (1987). Design and development of performance assessments. *Educational Measurement: Issues and Practices, 6*(3), 33-42.

Wright, B. D. (1996). Reliability and separation. *Rasch Measurement: Transactions of the Rasch Measurement SIG, 9*(4), 472.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago, IL: MESA Press.

Figure 1. Experimental Design for Day 1.

AM	Trainer 1 Group 1 (Raters 1, 2, 3, 4, 5) **R123VAX6/AM** Rating Scales a defined midpoint 3 hatchmarks Training anchors show endpoints and midpoint practice sets show full continuum	Trainer 2 Group 2 (Raters 6, 7, 8, 9, 10, 11) **R1VAX1/AM** Rating Scales no defined midpoint 5 hatchmarks Training anchors show endpoints only practice sets show full continuum
	LUNCH	LUNCH
PM	Trainer 2 Group 1 (Raters 1, 2, 3, 4, 5) **R1VAX1/PM** Rating Scales a defined midpoint no hatchmarks Training anchors show endpoints and midpoint practice sets show full continuum	Trainer 1 Group 2 (Raters 6, 7, 8, 9, 10, 11) **R123VAX6/PM** Rating Scales no defined midpoint no hatchmarks Training anchors show endpoints only practice sets show endpoints only

Figure 2. Experimental Design for Day 2.

AM	Trainer 1 Group 1 (Raters 1, 2, 8, 9, 10) **R23VAX5/AM** Rating Scales no defined midpoint 3 hatchmarks Training anchors show endpoints only practice sets show full continuum	Trainer 2 Group 2 (Raters 3, 4, 5, 6, 7, 11) **R23VAX7/AM** Rating Scales no defined midpoint no hatchmarks Training anchors show endpoints only practice sets show full continuum
	LUNCH	LUNCH
PM	Trainer 2 Group 1 (Raters 1, 2, 8, 9, 10) **R23VAX7/PM** Rating Scales a defined midpoint no hatchmarks Training anchors show endpoints and midpoint practice sets show full continuum	Trainer 1 Group 2 (Raters 3, 4, 5, 6, 7, 11) **R23VAX5/PM** Rating Scales a defined midpoint 5 hatchmarks Training anchors show all five scale points practice sets show full continuum

Table 1. Summary statistics for R1VAX1/AM

(Each scale has no defined midpoint and five hatchmarks.)

(Anchors show endpoints only; practice sets show full continuum)

	3-point scales	4-point scales	5-point scales	6-point scales	7-point scales	8-point scales	9-point scales	10-point scales
Student Separation	3.34	4.20	4.28	4.48	4.34	4.57	4.63	4.60
Student Separation Reliability	.83	.88	.89	.88	.89	.90	.89	.90
Intraclass Correlation	.76	.77	.79	.80	.81	.80	.82	.82
Most Probable from:								
Scale 1	3	4	5	6	7	8	8	7
Scale 2	3	4	5	6	6	7	6	7
Average Measure Difference:								
Scale 1	3	4	5	6	7	7	9	10
Scale 2	3	4	5	6	7	8	9	10

Table 2. Summary statistics for R1VAX1/PM

(Each scale has a defined midpoint and no hatchmarks.)

(Anchors show endpoints and midpoint; practice sets show full continuum)

	3-point scales	4-point scales	5-point scales	6-point scales	7-point scales	8-point scales	9-point scales	10-point scales
Student Separation	2.96	3.82	4.24	4.24	4.68	4.65	4.77	4.79
Student Separation Reliability	.83	.88	.91	.90	.92	.92	.92	.92
Intraclass Correlation	.80	.84	.85	.86	.87	.87	.88	.88
Most Probable from:								
Scale 1	3	4	5	6	7	7	9	9
Scale 2	3	4	5	6	7	7	8	9
Average Measure Difference:								
Scale 1	3	4	5	6	7	8	9	9
Scale 2	3	4	5	6	7	8	9	10

Table 3. Summary statistics for R123VAX6/AM

(Each scale has a defined midpoint and three hatchmarks.)

(Anchors show endpoints and midpoint; practice sets show full continuum)

	3-point scales	4-point scales	5-point scales	6-point scales	7-point scales	8-point scales	9-point scales	10-point scales
Student Separation	2.50	2.73	3.29	3.37	3.40	3.55	3.59	3.59
Student Separation Reliability	.78	.85	.88	.87	.89	.90	.90	.89
Intraclass Correlation	.76	.78	.80	.81	.81	.81	.82	.82
Most Probable from:								
Scale 1	3	4	5	5	5	7	6	6
Scale 2	3	4	5	4	5	5	6	7
Average Measure Difference:								
Scale 1	3	4	5	6	7	7	8	8
Scale 2	3	4	5	6	7	8	9	9

Table 4. Summary statistics for R123VAX6/PM

(Each scale has no defined midpoint and no hatchmarks.)

(Anchors show endpoints only; practice sets show endpoints only)

	3-point scales	4-point scales	5-point scales	6-point scales	7-point scales	8-point scales	9-point scales	10-point scales
Student Separation	2.35	3.34	3.47	3.69	3.88	3.87	3.92	3.97
Student Separation Reliability	.81	.89	.90	.91	.92	.91	.93	.93
Intraclass Correlation	.67	.70	.75	.74	.76	.75	.76	.77
Most Probable from:								
Scale 1	3	4	5	6	7	8	8	8
Scale 2	3	4	5	6	7	7	9	9
Average Measure Difference:								
Scale 1	3	4	5	6	7	8	9	8
Scale 2	3	4	5	6	7	8	9	10

BEST COPY AVAILABLE

Table 5. Summary statistics for R23VAX7/AM

(Each scale has no defined midpoint and no hatchmarks.)

(Anchors show endpoints only; practice sets show full continuum)

	3-point scales	4-point scales	5-point scales	6-point scales	7-point scales	8-point scales	9-point scales	10-point scales
Student Separation	3.06	3.66	4.18	4.07	4.26	4.34	4.43	4.36
Student Separation Reliability	.85	.85	.88	.88	.90	.89	.90	.89
Intraclass Correlation	.73	.83	.87	.84	.87	.88	.88	.89
Most Probable from:								
Scale 1	3	4	5	6	7	8	8	9
Scale 2	3	4	5	6	6	7	8	8
Average Measure Difference:								
Scale 1	3	4	5	6	7	8	9	10
Scale 2	3	4	5	6	7	8	9	10

Table 6. Summary statistics for R23VAX7/PM

(Each scale has a defined midpoint and no hatchmarks.)

(Anchors show endpoints and midpoint; practice sets show full continuum)

	3-point scales	4-point scales	5-point scales	6-point scales	7-point scales	8-point scales	9-point scales	10-point scales
Student Separation	2.79	3.50	3.55	3.79	3.85	3.89	4.03	4.03
Student Separation Reliability	.70	.73	.79	.79	.81	.80	.80	.82
Intraclass Correlation	.80	.86	.85	.88	.88	.89	.88	.89
Most Probable from:								
Scale 1	3	4	5	6	7	8	8	9
Scale 2	3	4	5	6	7	8	8	7
Average Measure Difference:								
Scale 1	3	4	5	6	7	8	9	9
Scale 2	3	4	5	6	7	8	9	10

Table 7. Summary statistics for R23VAX5/AM

(Each scale has no defined midpoint and three hatchmarks.)

(Anchors show endpoints only; practice sets show full continuum)

	3-point scales	4-point scales	5-point scales	6-point scales	7-point scales	8-point scales	9-point scales	10-point scales
Student Separation	2.87	3.51	3.88	3.91	4.19	4.06	4.19	4.18
Student Separation Reliability	.77	.80	.86	.87	.88	.88	.88	.88
Intraclass Correlation	.86	.89	.91	.91	.92	.92	.92	.93
Most Probable from:								
Scale 1	3	4	5	6	6	7	7	9
Scale 2	3	4	5	5	7	7	8	5
Average Measure Difference:								
Scale 1	3	4	5	6	7	8	9	10
Scale 2	3	4	5	6	7	8	9	10

Table 8. Summary statistics for R23VAX5/PM

(Each scale has a defined midpoint and five hatchmarks.)

(Anchors show all five scale points; practice sets show full continuum)

	3-point scales	4-point scales	5-point scales	6-point scales	7-point scales	8-point scales	9-point scales	10-point scales
Student Separation	2.94	3.37	3.81	3.87	4.08	4.19	4.28	4.23
Student Separation Reliability	.85	.86	.88	.89	.91	.91	.91	.90
Intraclass Correlation	.87	.88	.91	.92	.92	.92	.93	.93
Most Probable from:								
Scale 1	3	4	5	6	6	8	7	5
Scale 2	3	4	5	6	5	7	7	6
Average Measure Difference:								
Scale 1	3	4	5	6	7	8	9	10
Scale 2	3	4	5	6	7	8	9	10

Table 9. Number of Scale Points Supported by Each Scale		
	"Most Probable From" (Thresholds)	"Average Measure Differences"
R1VAX1/AM		
Scale 1	8	7
Scale 2	6	10
R1VAX1/PM		
Scale 1	7	9
Scale 2	7	10
R123VAX6/AM		
Scale 1	5	7
Scale 2	5	9
R23VAX6/PM		
Scale 1	8	9
Scale 2	7	10
R23VAX7/AM		
Scale 1	8	10
Scale 2	6	10
R23VAX7/PM		
Scale 1	8	9
Scale 2	8	10
R23VAX5/AM		
Scale 1	6	10
Scale 2	5	10
R23VAX5/PM		
Scale 1	6	10
Scale 2	6	10

Table 10. Comparisons of Student Separation Reliability Coefficients (ordered from high to low) for the Eight Blocks

(for 5-point scales)

Block	Student Separation Reliability	Defined Midpoint	# of Hatchmarks	Anchors Show:	Practice Sets Show:
R1VAX1/PM	.91	yes	0	endpoints and midpoint	full continuum
R123VAX6/PM	.90	no	0	endpoints only	endpoints only
R1VAX1/AM	.89	no	5	endpoints only	full continuum
R123VAX6/AM	.88	yes	3	endpoints and midpoint	full continuum
R23VAX7/AM	.88	no	0	endpoints only	full continuum
R23VAX5/PM	.88	yes	5	all five scale points	full continuum
R23VAX5/AM	.86	no	3	endpoints only	full continuum
R23VAX7/PM	.79	yes	0	endpoints and midpoint	full continuum

Table 11. Comparisons of Intraclass Correlation Coefficients (ordered from high to low) for the Eight Blocks

(for 5-point scales)

Block	Intraclass Correlation	Defined Midpoint	# of Hatchmarks	Anchors Show:	Practice Sets Show:
R23VAX5/PM	.91	yes	5	all five scale points	full continuum
R23VAX5/AM	.91	no	3	endpoints only	full continuum
R23VAX7/AM	.87	no	0	endpoints only	full continuum
R1VAX1/PM	.85	yes	0	endpoints and midpoint	full continuum
R23VAX7/PM	.85	yes	0	endpoints and midpoint	full continuum
R123VAX6/AM	.80	yes	3	endpoints and midpoint	full continuum
R1VAX1/AM	.79	no	5	endpoints only	full continuum
R123VAX6/PM	.75	no	0	endpoints only	endpoints only

Appendix A

Descriptive Graphic Rating Scales

Rater ID: _____ Block: **R123VAX6/AM** (Mural) Set: _____ Student ID: _____

Instructions: Rate the student's work using the two scales below. For each scale, place a vertical slash anywhere along the horizontal line.

1. How effectively did the student use elements (line, shape, or color) that draw attention from far away?

|_____|_____|

Not at all effectively
("Didn't do it")

- line, shape or color would not be seen well from far away
- wasn't thinking about distance
- lacks unity

Somewhat effectively
("Did it")

- line, shape or color would be seen moderately well from far away
- tentative
- too much negative space

Very effectively
("Did it well")

- line, shape, or color would be seen very distinctly from far away
- uses at least two of the elements well
- strong focal point or idea

2. How effectively did the student use the drawing space?

|_____|_____|

Not at all effectively
("Didn't do it")

- small random, or static images
- lots of negative space
- haphazard composition

Somewhat effectively
("Did it")

- attempt at visual organization
- uses more of the space, but still too much negative space
- tentative composition

Very effectively
("Did it well")

- visual organization emphasizes content
- effective use of negative space
- confident composition

Rater ID: _____ Block: **R123VAX6/PM** (Mural) Student ID: _____

Set: _____

Instructions: Rate the student's work using the two scales below. For each scale, place a vertical slash anywhere along the horizontal line.

1. How effectively did the student use elements (line, shape, or color) that draw attention from far away?

Not at all effectively
("Didn't do it")

- line, shape or color would not be seen well from far away
- wasn't thinking about distance
- lacks unity

Very effectively
("Did it well")

- line, shape, or color would be seen very distinctly from far away
- uses at least two of the elements well
- strong focal point or idea

2. How effectively did the student use the drawing space?

Not at all effectively
("Didn't do it")

- small random, or static images
- lots of negative space
- haphazard composition

Very effectively
("Did it well")

- visual organization emphasizes content
- effective use of negative space
- confident composition

Rater ID: _____ Block: **R1VAX1/AM** (Animal's Place) Student ID: _____

Set: _____

Instructions: Rate the student's work using the two scales below. For each scale, place a vertical slash anywhere along the horizontal line.

1. How effectively does the student deal with near and far shapes and overlap?

|_____|_____|_____|_____|

not at all effectively

- lacks presence of a foreground, middle ground and background
- no diminution of size as elements recede toward the background
- no use of overlap to indicate depth

very effectively

- contains a foreground, middle ground, and background
- elements show diminution of size as they recede toward the background
- use of overlap to indicate depth

2. How effectively does the overall composition integrate the way the animal has been placed? To what extent is the animal situated in the environment?

|_____|_____|_____|_____|

not at all effectively

- total disregard for setting
- free-floating image
- no ground line
- illustration is crowded out by other elements in the drawing
- no great loss to the drawing if the image were removed

very effectively

- creative placement of animal directly into landscape or environment
- strong sense of a ground line
- elements in the drawing sit well together

Note: If the student approaches the task from a decorative angle with an emphasis on pattern and repeated elements rather than the depiction of a three-dimensional or perspectival space, then consider the placement of the animal *within the overall design.*

Rater ID: _____ Block: **R1VAX1/PM** (Animal's Place) Student ID: _____

Set: _____

Instructions: Rate the student's work using the two scales below. For each scale, place a vertical slash anywhere along the horizontal line.

1. How effectively does the student deal with near and far shapes and overlap?

---

not at all effectively

- lacks presence of a foreground, middle ground and background
- no diminution of size as elements recede toward the background
- no use of overlap to indicate depth

somewhat effectively

- begins to suggest foreground, middle ground, and/or background
- there is a ground line. The animal may be placed on the ground with the sky above

very effectively

- contains a foreground, middle ground, and background
- elements show diminution of size as they recede toward the background
- use of overlap to indicate depth

2. How effectively does the overall composition integrate the way the animal has been placed? To what extent is the animal situated in the environment?

---

not at all effectively

- total disregard for setting
- free-floating image
- no ground line
- illustration is crowded out by other elements in the drawing
- no great loss to the drawing if the image were removed

somewhat effectively

- some regard for placement of cat, though nothing highly creative
- some sense of a ground line

very effectively

- creative placement of animal directly into landscape or environment
- strong sense of a ground line
- elements in the drawing sit well together

Note: If the student approaches the task from a decorative angle with an emphasis on pattern and repeated elements rather than the depiction of a three-dimensional or perspectival space, then consider the placement of the animal *within the overall design.*

Rater ID: _____ Block: **R23VAX5/AM** (Self portrait) Set: _____ Student ID: _____

Instructions: Rate the student's work using the two scales below. For each scale, place a vertical slash anywhere along the horizontal line.

1. How expressive is the student's self portrait?

|_____|_____|

Not at all expressive
("Didn't do it")

- does not show expressive qualities
- lacks feeling/mood

Very expressive
("Did it well")

- clearly shows expressive qualities
- shows feeling/mood

Note: Written responses should be used to 1) clarify the student's intentions, and/or 2) add insight into the expressive nature of the work. Absence of written response should NOT have a negative effect on scoring (i.e., if there is no written response OR the content of the written response is inconsistent with the image, then don't penalize the student).

2. How successfully did the student use materials and techniques to convey expression?

|_____|_____|

Not at all successfully
("Didn't do it")

- student was not able to convey expression with the materials selected
- student was not able to convey expression with the techniques selected
- no conscious relationship between materials, techniques and style

Very successfully
("Did it well")

- student was able to clearly convey expression with the materials selected
- student was able to clearly convey expression with the techniques selected
- shows strong relationship between materials, techniques, and style

Rater ID: _____ Block: **R23VAX5/PM** (Self portrait) Student ID: _____

Set: _____

Instructions: Rate the student's work using the two scales below. For each scale, place a vertical slash anywhere along the horizontal line.

1. How expressive is the student's self portrait?

|_____|_____|_____|_____|

Not at all expressive *("Didn't do it")*	***Somewhat expressive*** *("Did it")*	***Very expressive*** *("Did it well")*
• does not show expressive qualities • lacks feeling/mood	• suggests expressive qualities • suggests feeling/mood	• clearly shows expressive qualities • shows feeling/mood

Note: Written responses should be used to 1) clarify the student's intentions, and/or 2) add insight into the expressive nature of the work. Absence of written response should NOT have a negative effect on scoring (i.e., if there is no written response OR the content of the written response is inconsistent with the image, then don't penalize the student).

2. How successfully did the student use materials and techniques to convey expression?

|_____|_____|_____|_____|

Not at all successfully *("Didn't do it")*	***Somewhat successfully*** *("Did it")*	***Very successfully*** *("Did it well")*
• student was not able to convey expression with the materials selected • student was not able to convey expression with the techniques selected • no conscious relationship between materials, techniques and style	• student was able to <u>suggest</u> expression with the materials selected • student was able to <u>suggest</u> expression with the techniques selected • shows <u>some</u> awareness of the relationship between materials, techniques, and style	• student was able to <u>clearly</u> convey expression with the materials selected • student was able to <u>clearly</u> convey expression with the techniques selected • shows <u>strong</u> relationship between materials, techniques, and style

Rater ID: _____ Block: **R23VAX7/PM** (Miseries and Hope) Set: _____ Student ID: _____

Instructions: Rate the student's work using the two scales below. For each scale, place a vertical slash anywhere along the horizontal line.

1. To what extent are the ideas expressed original (i.e., inventive, ingenious, nonstandard) in concept and form?

---

not at all original

- cliché
- very vague
- somewhat irrelevant
- no differentiation

somewhat original

- perhaps cliché but expressed with some originality
- somewhat vague
- two strong panels and a nonexistent one fits this category

very original

- personal
- inventive
- unusual
- expert cliché
- differentiated

2. How appropriate are the forms to the ideas being expressed? Do the forms convey the intended meaning?

---

not appropriate

- very vague
- not readable as relevant to purpose
- primarily uses words

somewhat appropriate

- meaning is somewhat apparent but may be vague or unclear
- somewhat vague or unclear differentiation of forms

very appropriate

- style communicates meaning
- characteristics are fully realized
- panels are clearly different from one another
- forms and ideas are ingeniously connected

Rater ID: _____ Block: **R23VAX7/AM** (Miseries and Hope) Set: _____ Student ID: _____

Instructions: Rate the student's work using the two scales below. For each scale, place a vertical slash anywhere along the horizontal line.

1. To what extent are the ideas expressed original (i.e., inventive, ingenious, nonstandard) in concept and form?

---

not at all original

- cliché
- very vague
- somewhat irrelevant
- no differentiation

very original

- personal
- inventive
- unusual
- expert cliché
- differentiated

2. How appropriate are the forms to the ideas being expressed? Do the forms convey the intended meaning?

---

not appropriate

- very vague
- not readable as relevant to purpose
- primarily uses words

very appropriate

- style communicates meaning
- characteristics are fully realized
- panels are clearly different from one another
- forms and ideas are ingeniously connected